DISCOVERING PATTERNS OF INSURGENCY VIA SPATIO-TEMPORAL DATA MINING

James P. Rogers, James A. Shine, Shashi Shekhar, Mete Celik
U.S. Army ERDC, Topographic Engineering Center, VA, USA
{james.p.rogers.II,james.a.shine} @erdc.usace.army.mil
Department of Computer Science, University of Minnesota, MN, USA
{shekhar,mcelik}@cs.umn.edu

Abstract

The need to discover patterns in spatio-temporal (ST) data has driven much recent research in ST cooccurrence patterns. Early work focused on discovering spatial patterns such as co-location without examining the development of patterns over time or the temporal aspect of ST datasets. This paper describes a novel set of cooccurrence patterns called mixed-drove co-occurrence patterns (MDCOPs). They represent subsets of two or more different ST object-types whose instances are close to each other both spatially and temporally. However, mining MDCOPs is computationally very expensive due to complex interest measures, larger archived and historical datasets, and exponential growth in candidate patterns with the number of object-types. We propose a monotonic composite interest measure for discovering MDCOPs and two novel MDCOP mining algorithms. Analytical results show that the proposed algorithms are correct and complete. Experimental results also show that the proposed methods are computationally more efficient than naïve alternatives.

1. Introduction

The Army manages huge amounts of spatio-temporal (ST) data from a multitude of databases, and the volume of such data continues to expand as database archives grow and ST sensors increase in number and resolution. This data is stored in attributes, values and tables, with important relationships and patterns that are not known. Humans have limited capacities to find these patterns with analysis and knowledge discovery, making automated and semi-automated pattern analysis essential. Relationships between different ST events are vital to increasing knowledge and understanding of military challenges such as discovering and predicting insurgent attack patterns and tactics, and understanding the nature of asymmetric warfare. As a result, ST co-occurrence pattern mining has been the subject of much recent research.

Given a moving object database, our aim is to discover mixed-drove ST co-occurrence patterns (MDCOPs) representing subsets of different object-types whose instances are located close together in geographic space for a significant fraction of time. Unlike the objectives of some other ST co-occurrence pattern identification approaches where the pattern is the primary interest, in MDCOPs both the pattern and the nature of the different *object-types* are of interest.

An example of an MDCOP is in American football where two teams try to outscore each other by moving a football to the opponent's end of the field. Various complex interactions occur within and across teams to achieve this goal. These interactions involve intentional and accidental MDCOPs, the identification of which may help teams to study their opponent's tactics. Object-types may be defined by the roles of the offensive and defensive players, such as quarterback, running back, wide receiver, kicker, holder, linebacker, and cornerback. An MDCOP is a subset of these object-types (such as {kicker, holder} or {wide_receiver, cornerback}) that occur frequently. One example MDCOP involves offensive wide receivers, defensive linebackers, and defensive cornerbacks, and is called a Broken Blitz play. In this play, the objective of the offensive wide receivers is to outrun any linebackers and defensive backs and get behind them, catching an undefended pass and hopefully running untouched for a touchdown. This interaction creates an MDCOP between wide receivers and cornerbacks. An example Broken Blitz play is given in Fig. 1. It shows the positions of four offensive wide receivers (W.1, W.2, W.3, and W.4), two defensive cornerbacks (C.1 and C.2), two defensive linebackers (L.1 and L.2), and a quarterback (Q.1) in four time slots. The solid lines between the players show the neighboring players. The wide receivers W.1 and W.4 cross over each other and the wide receivers W.2 and W.3 run directly to the end zone of the field. Initially, the wide receivers W.1 and W.4 are co-located with cornerbacks C.1 and C.2 and the wide

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **DEC 2008**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **Discovering Patterns Of Insurgency Via Spatio-Temporal Data Mining**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army ERDC, Topographic Engineering Center, VA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited

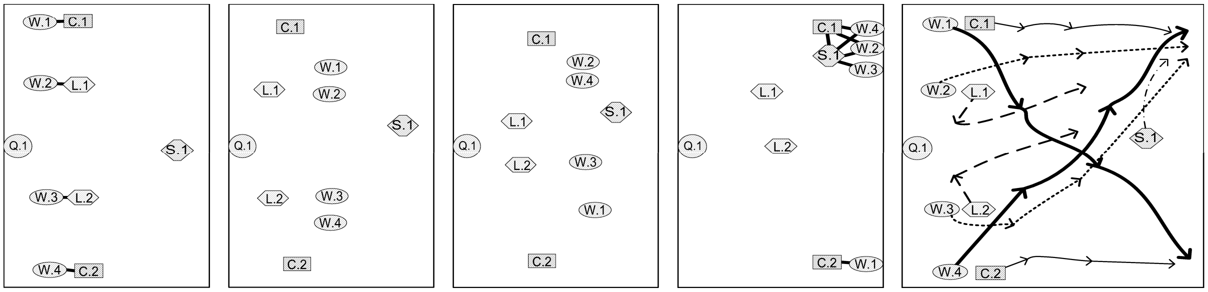
13. SUPPLEMENTARY NOTES

See also ADM002187. Proceedings of the Army Science Conference (26th) Held in Orlando, Florida on 1-4 December 2008

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**8**	



(a) time slot t=0 (b) time slot t=1 (c) time slot t=2 (d) time slot t=3 (e) coach sketch

Figure 1. An example Broken Blitz play in American football

receivers W.2 and W.3 are co-located with linebackers L.1 and L.2 at time slot t=0 (Fig. 1a). In time slot t=1, the four wide receivers begin to run, while the linebackers run towards the quarterback and the cornerbacks remain in their original position, possibly due to a fake handoff from the quarterback to the running back (Fig. 1b). In time slot t=2, the wide receivers W.1 and W.4 cross over each other and try to drift further away from their respective cornerbacks (Fig. 1c). When the quarterback shows signs of throwing the football, both cornerbacks and linebackers run to their respective wide receivers (Fig. 1d). The overall sketch of the game tactics can be seen in Fig. 1e. In this example, wide receivers and cornerbacks form an MDCOP since they are persistent over time and they occur in 2 out of 4 time slots. However, wide receivers and linebackers do not form an MDCOP due to the lack of temporal persistence.

Other applications for which discovering co-occurring patterns of specific combinations of object-types is important include battlefield planning and strategy, ecology (tracking species and pollutant movements), homeland defense (looking for significant "events"), and transportation (road and network planning) [8, 11].

However, discovering MDCOPs poses several non-trivial challenges. First, current interest measures (i.e. the spatial prevalence measure) are not sufficient to quantify such patterns, so new composite interest measures must be created and formalized [9]. Second, the set of candidate patterns grows exponentially with the number of object-types. Finally, since spatio-temporal datasets are huge, computationally efficient algorithms must be developed [16].

This paper focuses on MDCOPs (typed collections of moving objects) by extending interest measures for spatial co-location patterns given a user-defined participation index threshold [9]. The following issues are beyond the scope of this paper: (i) determining thresholds for MDCOP interest measures; (ii) similarity measures for tracking moving objects ; (iii) indexing and query processing issues related to mining objects; and (iv) discovering multisets (e.g.{A, A, B}).

2. Related Work

Data analysis can be broadly categorized into statistical approaches and data mining approaches. In statistical approaches, there are bodies of work in both spatial and temporal analysis. Spatial point patterns are often described by metrics such as the intensity function and Ripley's K [14, 15]. Other measures such as complete spatial randomness (CSR) and spatial covariance functions are used to describe the spatial relationships of adjacent areas and continuous variables as random fields [5]. Temporal patterns have been extensively studied in models such as moving averages, first and second order autoregression, integration, seasonality, and cointegration [18], [6]. There has also been some recent research in combining spatial and temporal analysis, such as Brix and Diggle's extended intensity function and the extended $K(r,t)$ function [1, 13]. Most attempts to combine the fields suffer from limitations such as the inability to model space-time interactions, assuming separability and independence between space and time [15]. Statistical research specifically focused on ST co-occurrence patterns and their possible interactions has been limited.

Previous data mining studies for mining ST co-occurrence patterns can be classified into two categories: mining of uniform groups of moving objects, and mining of mixed groups of moving objects.

To mine uniform groups of moving objects, the problems of discovering flock patterns [12, 7] and moving clusters [10] are defined. A flock pattern is a moving group of the same kind of objects, such as a sheep flock or a bird flock. Gudmundsson et al. proposed algorithms for detection of the flock pattern in ST datasets [7]. Kalnis et al. defined the problem of discovering moving clusters and proposed clustering-based methods to mine such patterns [10]. In this approach, if there is a large enough number of common objects between clusters in consecutive time slots, such clusters are called moving clusters. These methods do not take object-types into account, and thus are not effective for mining MDCOPs [4]. To mine mixed groups of moving objects, the problems of discovering collocation episodes [3] and topological patterns [17] are important. Both generalize co-location patterns [9] to the ST domain. A collocation episode is a sequence of co-location patterns with some common object-types across consecutive time slots. However, if there is no common object-type in consecutive time slots, the proposed approach will not identify any pattern. For example, if the window length is 2, the collocation episodes algorithm will not be able to find any pattern from the dataset given in Fig. 1. The algorithm tries to find co-location patterns that are persistent in 2 consecutive time slots, but there is no such pattern in the dataset because wide receivers and cornerbacks are forming co-locations in time slots t=0 and t=3 and wide receivers and linebackers are forming co-locations in time slots t=0. Thus, there may not be any co-location patterns and collocation episodes identified in the dataset.

A topological pattern [17] is a subset of object-types whose instances are close in space and time. An interest measure for a topological pattern $\{$A,B$\}$ (e.g. participation index or support) is a ST join of instances of A and instances of B [9]. This statistic may be high even if many instances of A and many instances of B are not spatially together for a moment in time. The semantics of topological patterns are not well-defined for moving objects. For example, this approach can not determine the fraction of time that a pattern occurs. This approach may not be able to tell in which time slots a pattern occurs, since there is no time slot notion. In the dataset given in Fig. 1, this approach will discover the two patterns of $\{$W,C$\}$, $\{$W,L$\}$, $\{$W,C$\}$, and $\{$W,C,S$\}$. Both patterns have the same support, but pattern $\{$W, C$\}$ occurs in 2 time slots out of 4 (a persistent pattern) and patterns $\{$W,L$\}$, $\{$W,C$\}$, and $\{$W,C,S$\}$ occur in 1 time slot out of 4 (a transient pattern) since tracks of objects are represented as ST instances. The persistent pattern $\{$W, C$\}$ occurs in time slots t=0 and t=3 and its instances $\{$W1, C1$\}$ and $\{$W4, C2$\}$ occur in time slot t=0 and $\{$W1, C2$\}$ and $\{$W4, C1$\}$ in time slot t=1. The transient pattern $\{$W, L$\}$ occurs in time slot t=0 and its instances $\{$W2, L1$\}$, $\{$W3, L1$\}$, $\{$W2, L2$\}$, and $\{$W3, L2$\}$ occur in time slot t=0.

In contrast, our proposed interest measure and algorithms will efficiently mine mixed groups of objects (e.g MDCOPs) which are close in space and persistent in time. Unlike the techniques just described, our approach discovers persistent patterns that co-occur in most but not all ST intervals; consecutive co-occurrences are not mandatory. For example, our approach will find the MDCOP $\{$wide_receiver, cornerback$\}$ pattern in Fig. 1, if the fraction of time slots where the pattern occurs over the total number of time slots is no less than the threshold 0.5, since instances are co-located in 2 time slots out of 4. It will reject the patterns $\{$W,L$\}$, $\{$W,C$\}$, and $\{$W,C,S$\}$ in Fig. 1 at the same threshold, since they are co-located on only 1 time slot out of 4.

3. Basic Concepts & Problem Definition

3.1 Spatial Prevalence Measure

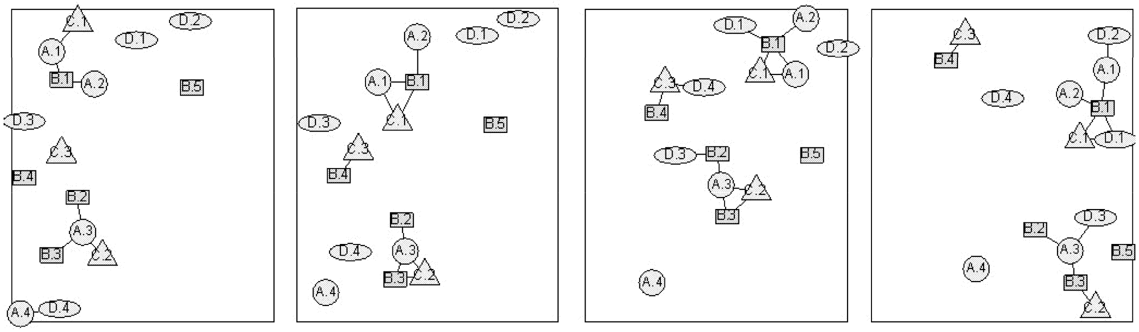
Spatial co-location mining algorithms are used to discover sets of mixed object-types that are frequently located together in a spatial framework for a given set of spatial object-types, their instances, and a spatial neighbor relationship *R* [9]. For example, in Fig. 2, in time slot t=0, $\{$A.1, C.1$\}$ is an instance of a co-location if the distance between the objects is no more than a given neighborhood distance threshold. In Fig. 2, the solid lines show the distance between the objects that satisfies the neighborhood distance threshold. The *participation index* is used to determine the strength of the co-location pattern. If the index is greater than or equal to a threshold [9], a co-location is called *spatial prevalent*. The participation index is defined as the minimum of the *participation ratios* (the fraction of the number of instances of object-types forming co-location instances to the total number of instances). For example, in Fig. 2, $\{$A, B$\}$ is a co-location in time slot t=0, and its instances are $\{$A.1, B.1$\}$, $\{$A.2, B.1$\}$, $\{$A.3, B.2$\}$, and $\{$A.3, B.3$\}$. In the dataset, object-type A has 4 instances and three of them (A.1, A.2, and A.3) are contributing to the co-location $\{$A, B$\}$, so the participation ratio of A is 3/4. The participation ratio of B is 3/5 since 3 out of 5 instances are contributing to the co-location $\{$A, B$\}$. The participation index of the co-location $\{$A, B$\}$ is the minimun of 4/5 and 3/5, or 3/5. It has been shown that the participation index is anti-monotone in the size of co-locations [9]. In other words, $participation\_index(P_j) \leq participation\_index(P_i)$ if P_i is a subset of P_j. In addition, the participation index has a spatial statistical interpretation as an upper bound on the $cross - K$ function [5].

3.2 Modeling MDCOPs

Given a set of spatio-temporal mixed object-types and a set of their instances with a neighborhood relation R, an MDCOP is a subset of spatio-temporal mixed object-types whose instances are neighbors in space and time.

Definition 3.1 *Given a spatio-temporal pattern and a set* TF *of time slots, such that* $TF = [T_0, ..., T_{n-1}]$, *the time prevalence or persistence measure of the pattern is the fraction of time slots where the pattern occurs over the total number of time slots.*

For example, in Fig. 2, the total number of time slots is 4 and pattern $\{$A, B$\}$ occurs in all 4 time slots, so its time prevalence index is 4/4. Pattern $\{$A, C$\}$ occurs in 3 time slots, namely, time slots t=0, t=1, and t=2, and its time prevalence index is 3/4 (Fig. 2b).



(a) An input spatio-temporal dataset

Mixed-Drove Spatio-Temporal Co-occurrence Patterns	Spatial prevalence index values				Time prevalence index values
	time slot 0	time slot 1	time slot 2	time slot 3	
A B	3/5	3/5	3/5	3/5	4/4
A C	2/4	2/4	2/4	0	3/4
B C	0	3/5	3/5	3/5	3/4
A B C	0	2/5	2/5	0	2/4

(b) A set of output mixed-drove spatio-temporal co-occurrence patterns

Figure 2. An input spatio-temporal dataset and a set of output MDCOPs

Definition 3.2 *Given a spatio-temporal dataset of mixed object-types ST, and a spatial prevalence threshold θ_p, the mixed-drove prevalence measure of pattern P_i is a composition of the spatial prevalence and the time prevalence measures as shown below.*

$$\mathrm{P}rob_{t_m \in TF}(s\_prev(P_i, time\_slot\ t_m) \geq \theta_p), \quad (1)$$

where $Prob$ stands for probability of overall prevalence time slots and s_prev stands for spatial prevalence, e.g., the participation index, described in Section 3.1.

Definition 3.3 *Given a spatio-temporal dataset of mixed object-types ST and a threshold pair (θ_p, θ_{time}), MDCOP P_i is a mixed-drove prevalent pattern if its mixed-drove prevalence measure satisfies the following.*

$$\mathrm{P}rob_{t_m \in TF}\left[s\_prev(P_i, time\_slot\ t_m) \geq \theta_p\right] \geq \theta_{time}, \quad (2)$$

where $Prob$ stands for probability of overall prevalence time slots, s_prev stands for spatial prevalence, θ_p is the spatial prevalence threshold, and θ_{time} is the time prevalence threshold.

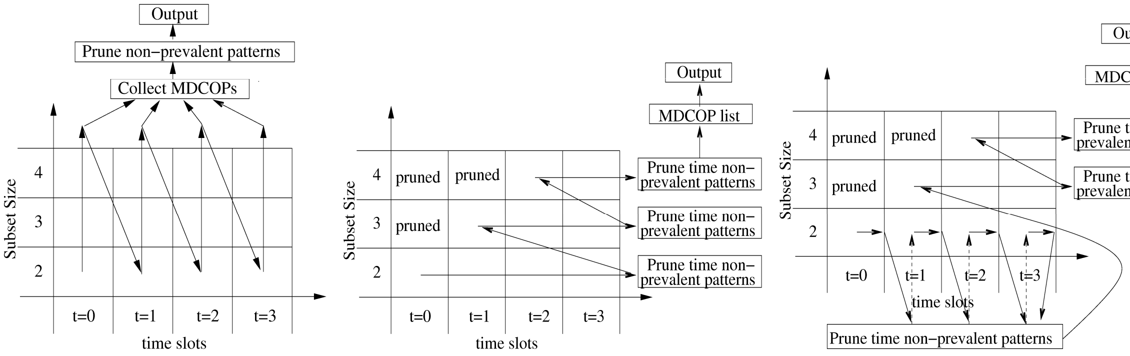
For example, in Fig. 2, {A, B} is an MDCOP because it is spatial prevalent in time slots t=0, t=1, t=2, and t=3 since its participation indices are no less than the given threshold 0.4 in these time slots, and is time prevalent since its time prevalence index of 1 is above the threshold 0.5. In contrast, {B, D} is not an MDCOP. Although it is spatial prevalent in time slot t=2, it is not time prevalent since its time prevalence index is no more than the given time prevalence threshold 0.5.

3.3 Problem statement

Given:
- A set P of Boolean spatio-temporal object-types over a common spatio-temporal framework STF.
- A neighbor relation R over locations.
- A spatial prevalence threshold, θ_P .
- A time prevalence threshold, θ_{time} .

Find: $\{P_i | P_i$ is a subset of P and P_i is a prevalent MDCOP as in Definition 3.3$\}$.
Objective: Minimize computation cost.
Constraints: To find a correct and complete set of MDCOPs.

Threshold values selected for MDCOP interest measures (the spatial and time prevalence measures) have important implications on the mining processes and results. Selection of a small interest measure threshold (close to 0) increases the algorithm's computational complexity and the number of generated prevalent patterns. This may cause generation of insignificant patterns. Selection of a large interest measure threshold (close to 1) decreases the computational complexity of the algorithms and the number of prevalent patterns. This may cause pruning of some of the significant patterns. Nevertheless the selection of interest measure threshold values is dependent on the application


Output
Prune non−prevalent patterns
Collect MDCOPs
Subset Size
4
3
2
t=0
t=1
t=2
t=3
time slots
Output
MDCOP list
Prune time non−prevalent patterns
pruned
Output
MDCOP list
Prune time non−prevalent patterns

(a) Naïve Approach

(b) MDCOP-Miner

(c) FastMDCOP-Miner

Figure 3. Comparison of Naïve Approach, MDCOP-Miner, and FastMDCOP-Miner

and/or purpose of the analysis.

4. Mining MDCOPs

In this section, we discuss a naïve approach and then propose two novel MDCOP mining algorithms – MDCOP-Miner and FastMDCOP-Miner – to mine MDCOPs.

4.1 Naïve approach

A naïve approach can use a spatial co-location mining algorithm for each time slot to find spatial prevalent co-locations and then apply a post-processing step to discover MDCOPs by checking their time prevalence. To mine co-locations, Huang, Shekhar and Xiong proposed a join-based approach, Yoo, Shekhar and Celik proposed a partial join-based approach and a join-less approach, and Zhang et al. proposed a multi-way spatial join-based approach [2, 9, 19, 20]. This study will be based on the join-based co-location algorithm proposed by Huang et al., but it is also possible to use other approaches. The naïve approach will generate size $k+1$ candidate co-locations for each time slot using spatial prevalent size k subclasses until there are no more candidates. After finding all size spatial prevalent co-locations in each time slot, a post-processing step can be used to discover MDCOPs by pruning out time non-prevalent co-locations. Even though this approach will prune out spatial non-prevalent co-locations early, it will not prune out time non-prevalent MDCOPs before the post-processing step (Fig. 3a). This leads to unnecessary computational cost.

4.2 MDCOP-Miner

To eliminate the drawbacks of the Naïve approach, we propose an MDCOP mining algorithm (MDCOP-Miner) which incorporates a time-prevalence based filtering step in each iteration. The algorithm will first discover all size k spatial prevalent MDCOPs, and then will apply a time-prevalence based filter to discover MDCOPS. Finally, the algorithm will generate size $k+1$ candidate MDCOPs using size k MDCOPs (Fig. 3b). The participation index is used as a spatial prevalence interest measure to check if the pattern is spatial prevalent at a time slot [9]. The time prevalence from Definition 3.1 is used as a time prevalence interest measure. Algorithm 1 gives the pseudo code of the both the MDCOP-Miner algorithm and the FastMDCOP-Miner discussed in the next section. The choice of the algorithm is provided by the user. The inputs are algorithm choice alg_choice with value *MDCOP-Miner*, a set of distinct spatial object-types E, a spatio-temporal dataset ST, a spatial neighborhood relationship R, and thresholds of interest measures, i.e. spatial prevalence and time prevalence; the output is a set of MDCOPs. In the algorithm, steps 1 include initialization of the parameters, steps 2 through 14 give an iterative process to mine MDCOPs, and step 15 gives a union of the results. Steps 2 through 14 continue until there are no candidate MDCOPs to be generated.

4.3 FastMDCOP-Miner

In this section, we discuss the FastMDCOP-Miner algorithm, which further improves on the computational efficiency of the MDCOP-Miner discussed in Section 4.2. As can be seen in Fig. 3b and in Algorithm 1, MDCOP-Miner waits to prune time non-prevalent patterns until all size k spatial prevalent patterns are generated for all time slots and then prunes time non-prevalent patterns to discover MDCOPs. However, we can further optimize the MDCOP-Miner by pruning time-non prevalent patterns at an earlier stage. We move ”prune non-prevalent patterns” between the time slots shown in Fig. 3c where the candidate size 2 pattern generation is illustrated. The

Algorithm 1:MDCOP-Miner and FastMDCOP-Miner

```
Inputs:
  $alg\_choice$:  MDCOP-Miner or FastMDCOP-Miner
  $E$:  a set of distinct spatial object-types
  $ST$:  a spatio-temporal dataset
      <object_type, object_id, x, y, time slot>
  $R$:  spatial neighborhood relationship
  $TF$:  a time slot frame $\{t_0, \ldots, t_{n-1}\}$
  $\theta_p$ :  a spatial prevalence threshold
  $\theta_{time}$ :  a time prevalence threshold
Output :    MDCOPs whose spatial prevalence
indices are no less than $\theta_p$, for time prevalence
indices are no less than $\theta_{time}$.
Variables:
  $k$:  co-occurrence size
  $t$:  time slots $(0, \ldots, n-1)$
  $T_k$:  set of instances of size $k$ co-occurrences
  $C_k$:  set of candidate size $k$ co-occurrences
  $SP_k$:  set of spatial prevalent size $k$
co-occurrences
  $TP_k$:  set of time prevalent size $k$
co-occurrences
  $MDP_k$:  set of mixed-drove size $k$
co-occurrences
Algorithm:
1) initialization :  $k = 1, C_k = E, MDP_k(0) = ST$
2) while (not empty $MDP_k$) {
3)   $C_{k+1}(0) = gen\_candidate\_co - occ(C_k, MDP_k)$
4)   for each time_slot $t$ in $(0, \ldots, n-1)$ {
5)     $T_{k+1}(t) = gen\_co\_occ\_inst(C_{k+1}(t), T_k(t), R)$
6)     $SP_{k+1}(t) = find\_spatial\_prev\_co\_occ(T_{k+1}(t), C_{k+1}(t), \theta_p)$
7)     If ($alg\_choice$ =="FastMDCOP-Miner") {
8)       $TP_{k+1}(t) = find\_time\_index(SP_{k+1}(t))$
9)       $MDP_{k+1}(t) = find\_time\_prev\_co\_occ(TP_{k+1}(t), \theta_{time})$
10)       $C_{k+1}(t) = MDP_{k+1}(t)$ } }
11)  If $alg\_choice$=="MDCOP-Miner" {
12)    $TP_{k+1} = find\_time\_index(SP_{k+1})$
13)    $MDP_{k+1} = find\_time\_prev\_co\_occ(TP_{k+1}, \theta_{time})$ }
14)  $k = k + 1$ }
15) return union $(MDP_2, \ldots, MDP_{k+1})$
```

pseudo-code of the FastMDCOP-Miner is also given in Algorithm 1. When the FastMDCOP-Miner is chosen, the algorithm will activate steps 8, 9, and 10 and deactivate steps 12 and 13. This will allow the algorithm to check the time prevalence of a pattern after every time slot is processed. The functions of the algorithm are as described in Section 4.2. In step 8, FastMDCOP-Miner checks whether the time prevalence indices of size k patterns (size 2 patterns in Fig. 3c) satisfy the time prevalence threshold before generating size k patterns for the next time slot. Early discovered time non-prevalent patterns are pruned in Step 9 and time prevalent patterns are used as candidate co-occurrences (Step 10) in the next time slot. For example, assume that there are 10 time slots and the time prevalence threshold is 0.5. In this case, a size k pattern should be present for at least 5 time slots to satisfy the threshold. If the time prevalence index of a pattern is 0 for the first (or any) 6 time slots, there is no need to generate it and check its prevalence for the rest of the time slots, since it will now be impossible for it to satisfy the given threshold regardless of the remaining results.

5. Experimental Evaluation

In this section, we present our experimental evaluations of several design decisions and workload parameters on our MDCOP mining algorithms. We used a real-world training dataset. Experiments were conducted on an IBM Netinfinity Linux Cluster, 2.6 GHz Intel Pentium 4 with 1.5 GB of RAM. We evaluated the behavior of the FastMDCOP-Miner, MDCOP-Miner and naïve approach to answer the following questions:

- What is the effect of the number of timeslots?
- What is the effect of the number of object-types?
- What is the effect of the spatial prevalence threshold?
- What is the effect of the time prevalence threshold?

Dataset: The real dataset contains the location and time information of moving objects (Figure 4). It includes 15 time snapshots and 22 distinct vehicle types and their instances. The minimum instance number is 2, the maximum instance number is 78, and the average number of instances is 19.

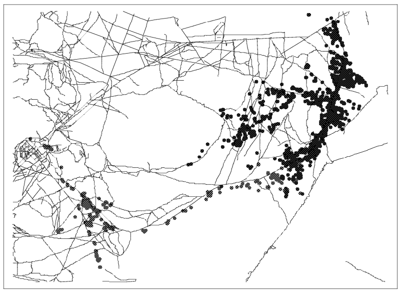
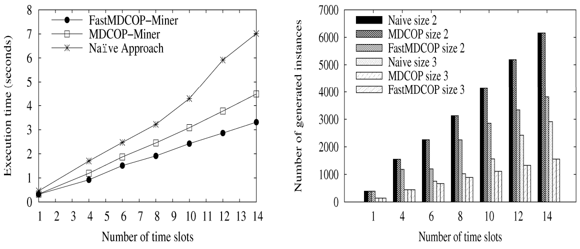


Figure 4. Real Dataset

1. Effect of Number of Time Slots: We evaluated the effect of the number of time slots on the execution time of the MDCOP algorithms using the real dataset. The participation index, time prevalence index, and distance were set at 0.2, 0.8, and 150m respectively. Experiments were run for a minimum of 1 time slot and a maximum of 14 time slots. Results showed that the FastMDCOP-Miner requires less execution time than the MDCOP-Miner and naïve approaches, since it prunes out time non-prevalent MDCOPs as early as possible (Fig. 5a). As the number of time slots increases, the ratio of the increase in execution time is smaller for FastMDCOP-Miner than for the other approaches. Fig. 5b shows the number of generated size 2 and size 3 instances for algorithms. The FastMDCOP-Miner generates fewer patterns. The MDCOP-Miner and naïve approaches generate the same number of size 2 instances.

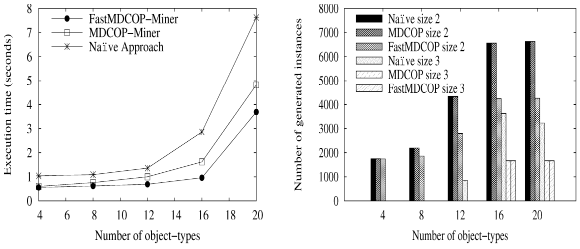
2. Effect of Number of Object-types: We evaluated the effect of the number of object-types on the execution time



(a) Execution time (b) Generated instances

Figure 5. Effect of number of time slots

of the algorithms using the real dataset. The participation index, time prevalence index, number of time slots and distance were set at 0.2, 0.8, 15, and 150m respectively. Results showed that the FastMDCOP-Miner outperforms the other approaches as the number of object-types increases (Fig. 6a-b). It is observed that the increase in execution time for the naïve approach is bigger than that of the MDCOP-Miner and the FastMDCOP-Miner as the number of object-types increases for datasets.

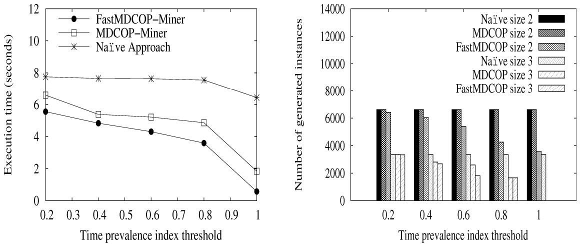


(a) Execution time (b) Generated instances

Figure 6. Effect of number of object-types

The execution times of the algorithms increases as the number of the object-types increases due to the increase in the number of join operations. The MDCOP-Miner and naïve approaches generate the same number of size 2 instances. In contrast, the FastMDCOP-Miner generates fewer size 2 instances (Fig. 6b).

3. Effect of the Time Prevalence Threshold: We evaluated the effect of the time prevalence threshold on the execution times of the MDCOP mining algorithms for the real dataset. The fixed parameters were participation index, number of time slots, and distance, and their values were 0.2, 15, and 150m respectively. For the naïve approach, the effective cost in execution time to generate spatial prevalent co-locations will be constant since it generates the same number of spatial prevalent patterns as the time prevalence index increases. In that case, the cost of the post-processing step will reflect the trend of the naïve approach. Experimental results show that the FastMDCOP-Miner is more computationally efficient than the other approaches (Fig. 7a-b). The execution times of the FastMDCOP-Miner and MDCOP-Miner decrease as the time prevalence threshold increases. It is also observed that the naïve approach is computationally more expensive as the time prevalence threshold decreases because of the increase in the number of MDCOPs to be discovered.



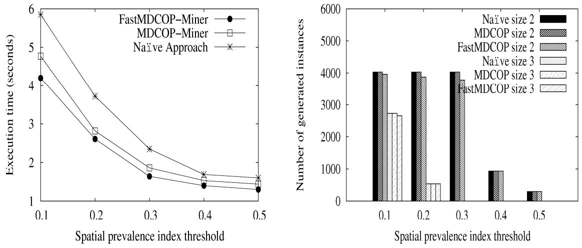
(a) Execution time (b) Generated instances

Figure 7. Effect of the time prevalence threshold

4. Effect of the Spatial Prevalence Threshold:We evaluated the effect of the spatial prevalence threshold on the execution times of the MDCOP algorithms. The fixed parameters were time prevalence index, number of time slots, and distance, with values of 0.2, 15, and 100m respectively. Fig. 8a shows the execution times of the algorithms and Fig. 8b shows the number of generated size 2 and 3 instances for the algorithms. FastMDCOP-Miner and MDCOP-Miner do not generate more than size 3 instances for a spatial prevalence threshold of greater than 0.2. The FastMDCOP-Miner outperforms the MDCOP-Miner and naïve approaches as the spatial prevalence threshold increases (Fig. 8a-b). The cost of the naïve approach is higher than that of the FastMDCOP-Miner and MDCOP-Miner for low values of the spatial prevalence threshold.

6. Conclusions and Future Work

We defined mixed-drove spatio-temporal co-occurrence patterns (MDCOPs) and the MDCOP mining problem and proposed a new monotonic composite interest measure which is the composition of distinct object-types, spatial prevalence measures, and time prevalence measures. We presented two novel and computationally efficient algorithms for mining these patterns: the MDCOP-Miner, and the FastMDCOP-Miner. We compared our algorithms with a naïve approach and showed their superiority in experi-



(a) Execution time

(b) Generated instances

Figure 8. Effect of the spatial prevalence threshold

ments using a real dataset to examine the effects of the number of time slots, the number of object-types, and the values of the spatial and time prevalance thresholds. The two new proposed algorithms are correct and complete in finding mixed-drove prevalent patterns.

For future work, we would like to explore the relationship between the proposed MDCOP interest measures and spatio-temporal statistical measures of interaction [2]. Another problem of interest is the characterization of the probability distribution of the proposed interest measure to help choose thresholds in the proposed measures. We plan to explore other potential interest measures for MDCOPs by evaluating similarity measures for tracks of moving objects. We plan to investigate new monotonic composite interest measures and develop other new computationally efficient algorithms for mining MDCOPs. We also hope to extend our algorithm to mine newly defined patterns in the literature such as leadership, convergence and query processing [12].

7. Acknowledgements

Some of the co-authors received support from USDOD and NSF. The authors would like to thank Kim Koffolt and Pradeep Mohan for their comments.

References

[1] A.Brix and P. Diggle. Spatio-temporal prediction for log-gaussian cox processes. *Journal of the Royal Statistical Society*, 63(10):823–841, 2001.

[2] S. Banerjee, B. P. Carlin, and A. E. Gelfrand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, 2003.

[3] H. Cao, N. Mamoulis, and D. W. Cheung. Discovery of collocation episodes in spatiotemporal data. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006)*, pages 823–827, Hong Kong, China, 2006.

[4] M. Celik, S. Shekhar, J. P. Rogers, and J. A. Shine. Mixed-drove spatio-temporal co-occurrence pattern mining. *IEEE Transacions on Knowledge and Data Engineering (Earlier version published in 6th IEEE International Conference on Data Mining 2006*, 2008.

[5] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, 1993.

[6] C. Granger. Time series analysis, cointegration, and applications. In *Nobel Prize lecture, Department of Economics, University of California, San Diego. Paper 2004-02.*, http://repositories.cdlib.org/ucsdecon/2004-02, 2004.

[7] J. Gudmundsson, M. v. Kreveld, and B. Speckmann. Efficient detection of motion patterns in spatio-temporal data sets. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems (ACM-GIS'04)*, pages 250–257, Washington DC, USA, 2004.

[8] R. Guting and M. Schneider. *Moving Object Databases*. Morgan Kaufmann, 2005.

[9] Y. Huang, S. Shekhar, and H. Xiong. Discovering co-location patterns from spatial datasets: A general approach. *IEEE Transactions on Knowledge and Data Engineering (TKDE) (Earlier version published in Symp. on Spatial and Temporal Databases 2001)*, 16(12):1472–1485, 2004.

[10] P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In *9th International Symp. on Spatial and Temporal Databases (SSTD)*, Angra dos Reis, Brazil, 2005.

[11] M. Koubarakis, T. Sellis, A. Frank, S. Grumbach, R. Guting, C. Jensen, N. Lorentzos, H. J. Schek, and M. Scholl. *Spatio-Temporal Databases: The Chorochronos Approach, LNCS 2520*, volume 9. Springer Verlag, 2003.

[12] P. Laube, M. v. Kreveld, and S. Imfeld. Finding remo - detecting relative motion patterns in geospatial lifelines. In *11th International Symp. on Spatial Data Handling*, pages 201–214. Springer Berlin Heidelberg, 2004.

[13] J. Ma, D. Zeng, and H. Chen. Spatial-temporal cross-correlation analysis. In *Proceedings of the 2006 IEEE International Conference on Intelligence and Security Informatics*, pages 542–547, San Diego, CA, 2006.

[14] B. Ripley. *Spatial Statistics*. Wiley, 1981.

[15] O. Schabenberger and C. Gotway. *Statistical Methods for Spatial Data Analysis*. Chapman and Hall, 2005.

[16] SSTDM06. First international workshop on spatial and spatio-temporal data mining. *In Conjunction with the 6th IEEE International Conference on Data Mining (ICDM 2006)*, 2006.

[17] J. Wang, W. Hsu, and M. L. Lee. A framework for mining topological patterns in spatio-temporal databases. In *ACM Fourteenth Conference on Information and Knowledge Management (CIKM'05)*, Bremen, Germany, 2005.

[18] W. W. S. Wei. *Time Series Analysis: Univariate and Multivariate Methods*. Addison Wesley, 2005.

[19] J. S. Yoo and S. Shekhar. A joinless approach for mining spatial colocation patterns. *IEEE Transactions on Knowledge and Data Engineering (TKDE) (partial results published in IEEE Int'l Conf on Data Mining 2005 and ACM Int'l Workshop on Geographic Inf. Systems 2005*, 18(10), 2006.

[20] X. Zhang, N. Mamoulis, D. W. L. Cheung, and Y. Shou. Fast mining of spatial collocations. In *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 384–393, Seatle, WA, 2004.